International Statistical Review

International Statistical Review (2021), 89, 1, 186–206 doi:10.1111/insr.12378

Random Effects Misspecification Can Have Severe Consequences for Random Effects Inference in Linear Mixed Models

Francis K. C. Hui[1], Samuel Müller[2] and Alan H. Welsh[1]

[1]*Research School of Finance, Actuarial Studies & Statistics, Australian National University, Canberra, Australia*
[2]*School of Mathematics and Statistics, The University of Sydney, Sydney, Australia*
E-mail: francis.hui@anu.edu.au

Summary

There has been considerable and controversial research over the past two decades into how successfully random effects misspecification in mixed models (i.e. assuming normality for the random effects when the true distribution is non-normal) can be diagnosed and what its impacts are on estimation and inference. However, much of this research has focused on fixed effects inference in generalised linear mixed models. In this article, motivated by the increasing number of applications of mixed models where interest is on the variance components, we study the effects of random effects misspecification on random effects inference in linear mixed models, for which there is considerably less literature. Our findings are surprising and contrary to general belief: for point estimation, maximum likelihood estimation of the variance components under misspecification is consistent, although in finite samples, both the bias and mean squared error can be substantial. For inference, we show through theory and simulation that under misspecification, standard likelihood ratio tests of truly non-zero variance components can suffer from severely inflated type I errors, and confidence intervals for the variance components can exhibit considerable under coverage. Furthermore, neither of these problems vanish asymptotically with increasing the number of clusters or cluster size. These results have major implications for random effects inference, especially if the true random effects distribution is heavier tailed than the normal. Fortunately, simple graphical and goodness-of-fit measures of the random effects predictions appear to have reasonable power at detecting misspecification. We apply linear mixed models to a survey of more than 4 000 high school students within 100 schools and analyse how mathematics achievement scores vary with student attributes and across different schools. The application demonstrates the sensitivity of mixed model inference to the true but unknown random effects distribution.

Key words: Fixed effects; hypothesis testing; maximum likelihood; prediction; robustness; variance components.

1 Introduction

Is it possible to detect misspecification of the random effects distribution in mixed models? What are the consequences of random effects misspecification on the estimation and inference of fixed and random effects? These questions have received considerable attention in the statistical literature over the past two decades and deservedly so given the increasing use of

mixed models in all areas of applied statistics (e.g. Bolker *et al.*, 2009; Tashakkori & Teddlie, 2010). Moreover, there is an increasing push towards hierarchical models incorporating random effects at multiple levels to reflect various scientific processes (Cressie *et al.*, 2009). Consequently, identifying whether we can assess if the assumed random effects distribution is appropriate or not and understanding how random effects misspecification impacts various aspects of inference are issues of major relevance to modern statistics.

There is an extensive and controversial literature on the severity of random effects misspecification in mixed models, which can be roughly divided into cases where the evidence is in agreement and cases where there is conflicting evidence. An excellent review was conducted by McCulloch and Neuhaus (2011a), although it should be acknowledged that their review was influenced by their own approach to the topic (see also Grilli & Rampichini, 2015). It is not the aim of this article to present a comprehensive review of this literature, but we do summarise the areas of agreement and disagreement about misspecifying the random effects distribution.

Overall, there are three main aspects of mixed models inference where the literature is in agreement. First, for estimation and inference of within-cluster coefficients, for example, time effects, and covariates that are included as a fixed effect only, there is coherent evidence that misspecification incurs relatively little bias, and type I errors and coverage probabilities remain close to their nominal levels (e.g. Zhang & Davidian, 2001; Litire *et al.*, 2007, 2008; Neuhaus *et al.*, 2013). Second, for estimation and inference of the fixed intercept in all mixed models, there is a general consensus that this is not robust to random effects misspecification. When the random intercept and/or other random effects distributions are misspecified, moderate-to-large bias can arise when estimating the fixed intercept (Heagerty & Kurland, 2001; Neuhaus *et al.*, 2013), and the type I error of Wald tests can be inflated above the nominal level when random effects are misspecified (Litire *et al.*, 2007, 2008). Third, for prediction of random effects, there is largely agreement that using the best predictor, that is, posterior mean, assuming normality of the random effects leads to reasonable performance [as assessed by the mean squared error (MSE) of prediction] even if the true random effects distribution is not normal (McCulloch & Neuhaus, 2011b). On the other hand, for assessing the shape of random effects distribution itself, the predictions are highly sensitive to misspecification of the assumed random effects distribution (McCulloch & Neuhaus, 2011b; Zhang & Davidian, 2001). It should be noted however that, compared with points 1 and 2 concerning the fixed effects, there is comparably less literature on how misspecification affects random effects prediction.

There are two main aspects of inference where the literature is not in agreement concerning the impact of random effects misspecification. First, for estimation and inference on between-cluster fixed effects, for example, treatment effects, and covariates that are included as both fixed and random effects, Litiere *et al*. (2007, 2008, 2011) in the context of logistic mixed models found moderate biases in the point estimates and large variations in the power to detect significant effects via hypothesis testing when the random effects were misspecified. For linear mixed models (LMMs), Zhang and Davidian (2001) similarly found that the standard normal random effects assumption, if incorrect, could lead to substantial losses in power when testing the significance of between-cluster coefficients. Other articles that have found evidence that misspecification can have major consequences on estimation and inference for between-cluster effects, and for fixed effect coefficients for covariates entered as both fixed and random effects, include Magder and Zeger (1996), Heagerty and Kurland (2001) and Agresti *et al*. (2004). These results conflict with those of Neuhaus *et al*. (2011, 2013), who presented asymptotic and empirical arguments for generalised LMMs to show that for between-cluster effects, there is generally negligible bias in point estimation, little to no efficiency loss in hypothesis testing, and that the coverage probability of Wald confidence intervals is close to the nominal level. Both McCulloch and Neuhaus (2011a) and Neuhaus *et al*. (2011) proceed to argue that the

results of Litire *et al*. (2007, 2008) do not directly address the question of misspecification for various reasons, while Neuhaus *et al.* (2013) also pointed out that the severe bias and issues with type I error power occurred only for 'very extreme, unrealistic random effect distributions when the true random effects distribution is highly discrete and/or the true variance components estimates are very large'. Second, for estimation and inference of variance components, Litire *et al.* (2008) presented evidence that variance components can be severely biased where the true random effects distribution is not normal. This occurred even when the true variance components were relatively small, although the bias did worsen when the true magnitude of the variance increased. This contrasts with the review of McCulloch and Neuhaus (2011a), who concluded that while not definitive, most of the evidence so far presents a case that the estimation of variance components is robust to random effects misspecification. It should be stressed however that, like random effects prediction, there is comparably less literature on the impact of random effects misspecification on this aspect of inference.

From the summary of the current literature earlier, one glaring point (also noted by McCulloch & Neuhaus, 2011a) is that much of the focus and debate has been on how random effects misspecification affects fixed effects estimation and inference. A major reason for this is because in most applications of mixed models in epidemiology and biostatistics, the focus is on the fixed effect parameters, while the random effects are there to account for between subject heterogeneity and are 'not often of interest' (Neuhaus *et al.*, 2013). While understandable, we believe that this viewpoint is too narrow, as in many other disciplines, inference on the random effects is indeed of scientific interest. For instance, Jiang and Lahiri (2006) and Schielzeth and Nakagawa (2013) provide general discussions of the utility of random effects inference in the context of ecology and small area estimation, respectively, and we refer to Martin *et al.* (2011), Ives and Helmus (2011), Nakagawa and Schielzeth (2013), Li *et al.* (2017), Smith *et al.* (2015) among others for specific examples in genetic, phylogeny and trait-based mixed model studies where the focus of the analysis is on assessing the significance of and interpreting the variance components or some function of it. From a statistical perspective, there are a variety of methods that have been developed that focus in some way on the random effects as the quantity of interest. Two notable examples are hierarchical and double generalised linear models (Lee *et al.*, 2006), which broaden generalised LMMs to directly model the dispersion and variance components as a function of covariates, and latent variable models where the multivariate random effects (latent variables) are allowed to vary as a function of covariates or are used directly for ordination (e.g. Hui, 2017; Woodard *et al.*, 2013).

Motivated by the comparative lack of research and the increasingly wide application of random effects inference in many scientific disciplines, we set out to provide greater insight into how misspecification of the random effects distribution impacts random effects estimation and inference in LMMs. Theoretically, we explore how the asymptotic properties of the standard likelihood ratio test differ between fixed effects coefficients and variance components under random effects misspecification. We complement this with a simulation study to explore the finite sample behaviour of estimation and inference under random effects misspecification. Finally, we applied LMMs to the 1982 High School and Beyond (HSB, Zwick & Sklar, 2005) survey, analysing how mathematics achievement scores vary with student gender, race and social economic status, as well as between schools. By modifying the dataset to vary the random effects distribution, we show that random effects inference, and to a lesser extent fixed effects inference, is sensitive to the true but unknown random effects distribution.

The focus of this article differs from recent literature in two key ways. First, attention of late has largely turned towards studying misspecification in generalised and non-LMMs (as exemplified by McCulloch & Neuhaus, 2011a), while we choose to focus our simulation on LMMs. Early seminal work by Butler and Louis (1992) and Verbeke and Lesaffre (1997) showed that,

even if incorrect, assuming normality for the random effects distribution in an LMM still produces maximum likelihood estimates of the fixed effects and variance components that are consistent and asymptotically normally distribution (with the sandwich covariance matrix). This result seems to have created a general belief that the consequences of random effects misspecification on *all* aspects of estimation and inference in LMMs is much less severe. We believe that this is a misguided belief, as asymptotic correctness does not necessarily equate to strong finite sample performance. Indeed, in the simulations accompanying their theory, Verbeke and and Lesaffre (1997) acknowledge that the rate of convergence of the MSE to zero for the point estimates 'heavily depends on the correct random effects distribution', especially for estimation of the random effects covariance matrix. Studying the theoretical and empirical impact of random effects misspecification on random effects inference in LMMs is thus an important contribution that we make to this underdeveloped portion of the literature. Second, while the majority of recent simulations in the literature have focused on the consequences of misspecification, we assess both this and the issue of whether we can *in some simple manner* even detect random effects misspecification. While there has been considerable research into methods for diagnosing random effects misspecification (e.g. Schützenmeister & Piepho, 2012; Verbeke & Molenberghs, 2013; Drikvandi *et al.*, 2017; Efendi *et al.*, 2017), most have garnered little attention with applied researchers favouring simpler approaches such as normal quantile plots accompanied by a test of normality (Pinheiro & Bates, 2006; Gaecki & Burzykowski, 2013). Therefore, we study whether we can detect violation of the normal random effects distribution assumption in a simple manner.

The three main findings of our theoretical and simulation study on random effects inference are as follows:

- Simple graphical exploration and goodness-of-fit measures based on the random effects predictions had fair-to-good power to detect departures from the assumed normal random effects distributions in LMMs, although there is some sensitivity to the shape of the true random effects distribution and the true value of the variance component. Increasing the number of clusters and/or cluster size improved the ability to diagnose random effects misspecification.
- Incorrectly assuming a normal distribution for the random effects can have severe consequences for random effects inference. Estimates of the variance components can be strongly biased, and this leads to potentially large MSEs relative to fixed effects inference. Type I errors for a standard likelihood ratio test that the variance component is truly non-zero were very sensitive to the true random effects distribution and were substantially inflated when the true distribution is heavier tailed than the normal. Analogously, profile likelihood confidence intervals for the variance components exhibited severe under coverage under random effects misspecification, with performance being especially poor for heavy-tailed distributions.
- For both random effects point estimation and hypothesis testing, increasing the cluster size had little impact on alleviating these consequences of misspecifying the random effects distribution. Increasing the number of clusters does improve point estimation but is *not* a remedy for hypothesis testing. Indeed, we show theoretically and empirically that the problem of random effects misspecification does not asymptotically vanish when it comes to testing truly non-zero variance components using standard likelihood ratio tests.

It is important to point out that, for significance testing, we focused on the theoretical and empirical behaviour of the likelihood ratio statistic when testing truly *non-zero* variance components. Almost all hypothesis testing of variance components however is at the zero boundary, for which it has been shown that type I errors of the likelihood ratio test for assessing whether variance components equal zero *are* robust to random effects misspecification. This however

does not mean our findings are mute: as with all parameters in a regression model, we are often interested in obtaining a confidence interval for the variance component or functions of it, for example, intra-class correlations. Indeed, the simulation results we conducted on coverage probability of confidence intervals for variance components show precisely their sensitivity to random effects misspecification, as they exhibit severe under coverage when the true random effects distribution is heavier tailed than the normal. Importantly, such under coverage does not asymptotically vanish.

2 Notation

We base our development around the independent cluster LMM, often used in longitudinal and single-level nested data, and consider random effects misspecification in the sense that the true random effects distribution is not normal, but we assume normality. Let y_{ij} denote measurement $j = 1, \ldots, m_i$ for cluster $i = 1, \ldots, n$. In conjunction, we observe a p_f-vector of fixed effects covariates $\boldsymbol{x}_{ij}$ and a p_r-vector of random effects covariates $\boldsymbol{z}_{ij}$. Typically, the first element in $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ is equal to one, corresponding to a fixed and random intercept, respectively. For each cluster $i = 1, \ldots, n$, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})$ denote the vector of responses, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1} \ldots \boldsymbol{x}_{im_i})^\top$ denote the $m_i \times p_f$ model matrix of fixed effects covariates and $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1} \ldots \boldsymbol{z}_{im_i})^\top$ denote the $m_i \times p_r$ model matrix of random effects covariates. Conditional on a p_r-vector of random effects $\boldsymbol{b}_i$, the response vector $\boldsymbol{y}_i$ is multivariate normally distributed with mean vector $\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\Lambda}\boldsymbol{b}_i$, where $\boldsymbol{\beta}$ denotes the vector of fixed effects coefficients, and covariance matrix $\sigma^2\boldsymbol{I}_{m_i}$, where $\boldsymbol{I}_{m_i}$ is a diagonal matrix of dimension m_i. The quantity $\boldsymbol{\Lambda}$ is defined as the Cholesky decomposition of the random effects covariance matrix. Furthermore, the vector $\boldsymbol{b}_i$ is drawn from a standardised random effects distribution, satisfying $\mathrm{E}(\boldsymbol{b}_i) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{b}_i) = \boldsymbol{I}_{p_r}$. This implies $\mathrm{Cov}(\boldsymbol{\Lambda}\boldsymbol{b}_i) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$.

Let $\boldsymbol{\Psi} = (\boldsymbol{\beta}^\top, \mathrm{vech}(\boldsymbol{\Lambda})^\top, \sigma^2)^\top$. For the LMM earlier, the marginal log likelihood is given by $\ell(\boldsymbol{\Psi}) = \sum_{i=1}^n \log\left(\int f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\beta}, \mathrm{vech}(\boldsymbol{\Lambda}), \sigma^2) f(\boldsymbol{b}_i) d\boldsymbol{b}_i\right)$, where $f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\beta}, \mathrm{vech}(\boldsymbol{\Lambda}), \sigma^2) = \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\Lambda}\boldsymbol{b}_i, \sigma^2\boldsymbol{I}_{m_i})$ is the conditional density of the responses and $f(\boldsymbol{b}_i)$ is the density of the assumed standardised random effects distribution. When the density of the random effects is also assumed to be normally distributed, then $f(\boldsymbol{b}_i) = \mathcal{N}_{p_r}(\boldsymbol{0}, \boldsymbol{I}_{p_r})$ and the integration can be performed analytically such that

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log\{\det(\boldsymbol{V}_i)\} - \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \right), \tag{1}$$

where $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{Z}_i^\top + \sigma^2\boldsymbol{I}_{m_i}$. When the random effects $\boldsymbol{b}_i$ follow some other non-normal standardised distribution, then the marginal log likelihood involves an intractable integral and numerical methods are typically applied to overcome this. Importantly however, because of the identity link function and standardised random effects distribution, then even under random effects misspecification, it holds that $\mathrm{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ and $\mathrm{Cov}(\boldsymbol{y}_i) = \boldsymbol{V}_i$.

3 Asymptotic Properties of Likelihood Ratio Tests

We establish some results concerning the use of likelihood ratio tests for testing fixed effects coefficients and variance components under random effects misspecification. The results build on the more general developments of Kent (1982), the relevance of which seems to have been somewhat overlooked in the literature on misspecified mixed models. Let $\hat{\boldsymbol{\Psi}} =$

$(\hat{\boldsymbol{\beta}}^\top, \text{vec}(\hat{\boldsymbol{\Lambda}})^\top, \hat{\sigma}^2)^\top$ denote the maximum likelihood estimates of the LMM obtained from maximising (1). Also, let $\boldsymbol{\Psi}^* = (\boldsymbol{\beta}^{*\top}, \text{vech}(\boldsymbol{\Lambda}^*)^\top, (\sigma^*)^2)^\top$ be the vector of true parameter values. In this article, we assume that none of the variance components are truly equal to zero, that is, $\boldsymbol{\Lambda}^*(\boldsymbol{\Lambda}^*)^\top$ is a positive definite covariance matrix. There has been extensive research conducted into random effects testing on the boundary of the parameter space (see, for instance, the work by Stram & Lee, 1994). Our focus here is on the impact of random effects misspecification on standard likelihood ratio tests, so we choose to avoid this added layer of complexity and concentrate on testing away from the boundary; see also our discussion at the end of Section 1.

To begin, note that in the presence of random effects misspecification and under general regularity conditions, the maximum likelihood estimates for LMMs have been proven to be strongly consistent, $\hat{\boldsymbol{\Psi}} \xrightarrow{\text{a.s}} \boldsymbol{\Psi}^*$ (see Verbeke & Lesaffre, 1997, for details). Next, let $\boldsymbol{\beta}_s$ denote any subvector of $\boldsymbol{\beta}$, and suppose we want to test the null hypothesis $H_0 : \boldsymbol{\beta}_s = \boldsymbol{\beta}_s^*$, where $\boldsymbol{\beta}_s^*$ is the corresponding subvector of $\boldsymbol{\beta}^*$. To do this, we consider the likelihood ratio statistic $\text{D}_{\boldsymbol{\beta}} = -2\{\ell(\tilde{\boldsymbol{\Psi}}) - \ell(\hat{\boldsymbol{\Psi}})\}$, where $\tilde{\boldsymbol{\Psi}}$ denotes the maximum likelihood estimate under H_0. Then we have the following result.

Result 1. *Under misspecification of the random effects distribution, the standard likelihood ratio test for fixed effects, where $D_{\boldsymbol{\beta}}$ is compared against a chi-squared distribution with degree of freedom equal to* $\dim(\boldsymbol{\beta}_s)$*, achieves the nominal significance level as $n \to \infty$. That is,*

$$P(D_{\boldsymbol{\beta}} > \chi^2_{1-\alpha}\{\dim(\boldsymbol{\beta}_s)\}|H_0) \to \alpha,$$

where $\chi^2_{1-\alpha}\{\dim(\boldsymbol{\beta}_s)\}$ denotes the $(1-\alpha)$ quantile of a chi-squared distribution with $\dim(\boldsymbol{\beta}_s)$ *degrees of freedom.*

The earlier result ensures the asymptotically validity of the commonly applied likelihood ratio test for testing any set of fixed effects in LMMs, even under random effects misspecification. Of course, asymptotic validity need not imply good finite sample performance, although both the previous literature (see the discussion in Section 1) and the simulation results in Section 5 show that empirically, the likelihood ratio test generally maintains a type I error close to the nominal level for a broad range of true random effects distributions. Next, we provide a result for the power of the likelihood ratio test under random effects misspecification.

Result 2. *Under misspecification of the random effects distribution, the power of the standard likelihood ratio test for fixed effects under the local alternative hypothesis $H_A : \boldsymbol{\beta}_s = \boldsymbol{\beta}_s^* + n^{-1/2}\boldsymbol{\delta}_s$, where $\|\boldsymbol{\delta}_s\|$ is a positive constant, is given by*

$$P(T(\boldsymbol{\delta}_s) > \chi^2_{1-\alpha}\{\dim(\boldsymbol{\beta}_s)\}|H_A),$$

as $n \to \infty$, where $T(\boldsymbol{\delta})$ is a non-central chi-squared distribution with degree of freedom equal to $\dim(\boldsymbol{\beta}_s)$ *and non-centrality parameter $K(\boldsymbol{\delta}_s)$ whose form is given in the Supporting Information.*

The earlier result is analogous to the more general result relating to the asymptotic alternative distribution of the likelihood ratio test (Van der Vaart, 2000), with the added complication being that $K(\boldsymbol{\delta}_s)$ is more complex to calculate because of random effects misspecification. More broadly, under non-local alternative hypotheses, it can be straightforwardly shown that the likelihood ratio test does not possess a non-degenerate limit distribution. That is, $\text{D}_{\boldsymbol{\beta}} \xrightarrow{p} \infty$, and hence, the power of the test tends to 1 as $n \to \infty$.

Consider now testing a set of variance components or equivalently testing a set of standard deviation components. For ease of presentation, we consider the case where $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{p_r})^\top$, and let $\boldsymbol{\lambda}_s$ denote any subvector of $\boldsymbol{\lambda}$. Suppose we test the null hypothesis $H_0 : \boldsymbol{\lambda}_s = \boldsymbol{\lambda}_s^*$, where $\boldsymbol{\lambda}_s^*$ is the corresponding subvector of $\boldsymbol{\lambda}^*$ and the elements of $\boldsymbol{\lambda}^*$ are positive. Then for the likelihood ratio statistic $\text{D}_{\boldsymbol{\lambda}} = -2\{\ell(\tilde{\boldsymbol{\Psi}}) - \ell(\hat{\boldsymbol{\Psi}})\}$, where $\tilde{\boldsymbol{\Psi}}$ denotes the maximum likelihood estimate under H_0, we have the following result.

Result 3. *Under misspecification of the random effects distribution, the standard likelihood ratio test for testing non-zero variance components, where $D_{\boldsymbol{\lambda}}$ is compared against a chi-squared distribution with degree of freedom equal to* $\dim(\boldsymbol{\lambda}_s)$*, is not guaranteed to achieve the nominated significance level as $n \to \infty$. That is,*

$$P(D_{\boldsymbol{\lambda}} > \chi^2_{1-\alpha}\{\dim(\boldsymbol{\lambda}_s)\}|H_0) \to \kappa,$$

for some $\kappa \in [0, 1]$ with κ not necessarily equal to α. Moreover, under the null hypothesis $H_0 : \boldsymbol{\lambda}_s = \boldsymbol{\lambda}_s^*$ with $\min\{\boldsymbol{\lambda}_s^*\} > 0$*, the statistic $D_{\boldsymbol{\lambda}}$ does not asymptotically follow a chi-squared distribution with* $\dim(\boldsymbol{\lambda}_s)$ *degrees of freedom.*

Result 3 implies that, when the true standard deviation components are away from the zero boundary, the standard likelihood ratio test for assessing standard deviation (and analogously variance) components is not guaranteed to be an asymptotically valid test. The value of κ, and moreover the asymptotic null distribution of $D_{\boldsymbol{\lambda}}$, depends on the shape of the true random effects distribution. Intuitively, Result 3 holds because the behaviour of the likelihood ratio test when testing $\boldsymbol{\lambda}$ relies on specific conditions involving the higher moments of the marginal distribution of $\boldsymbol{Y}$, that is, skewness and kurtosis, and generally, these conditions are not satisfied when the true random effects distribution is not normal.

Unlike the case of fixed effects hypothesis testing, with Result 3 in mind, there is little motivation in developing theoretical results on the power of the likelihood ratio test for testing non-zero variance components, and so we do not pursue such developments.

Although we have focused on significance testing, the results earlier may be extended to cover standard Wald and score tests, as well as various types of confidence intervals (see also Verbeke & Lesaffre, 1997). This is important because when the variance components are believed to be truly non-zero, then random effects inference is usually focused more on confidence intervals and uncertainty quantification rather than significance testing; see also the discussion at the end of Section 1. As we will see in numerical studies later on, the two are closely linked, and good (poor) performance of hypothesis testing tends to go hand in hand with good (poor) performance of the confidence intervals for the same parameters.

4 Simulation Study

We adopted the simulation design used in Neuhaus *et al.* (2013) and considered an independent cluster LMM with a random intercept and slope. We set $p_f = 3$ with the first column of $\boldsymbol{X}_i$ a column of ones representing a fixed intercept, the second column consisting of m_i ascending elements equally spaced on the interval -1 to $+1$ representing a centred within-cluster effect and the elements of the third column of $\boldsymbol{X}_i$ simulated randomly from a uniform distribution U[-1, 1]. Next, we set $p_r = 2$ with the two columns of $\boldsymbol{Z}_i$ equal to the first two columns of $\boldsymbol{X}_i$, reflecting a random intercept and slope. Furthermore, we let $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^\top$, such that λ_1 and λ_2 are the standard deviation components of the LMM. Finally, the two elements $\boldsymbol{b}_i = (b_{i0}, b_{i1})^\top$ are simulated independently from a variety of standardised random effects distributions.

As true parameter values, we set $\boldsymbol{\beta} = (\beta_0 = 2, \beta_1 = 1, \beta_2 = 1)^\top$ and $\sigma^2 = 1$ for all simulations and varied the true standard deviation components as $\lambda_1 = \lambda_2 = c_0 = \{0.5, 1, \sqrt{2}, 2, 3\}$. We considered combinations of $n = \{50, 100\}$ and equal cluster sizes $m_1 = \ldots = m_n = m = \{10, 20\}$ to explore how the ability to diagnose and the consequences of random effects misspecification are impacted by increasing number of clusters and/or cluster size.

For the true random effects, we simulated b_{i0} and b_{i1} independently from the following distributions: (1) a standard normal distribution (no misspecification); (2) a t-distribution with 3 degrees freedom, $t(3)$; (3) a Tukey gh distribution with $(g = 0.25, h = 0.05)$; (4) a Tukey gh distribution with $(g = 0.5, h = 0.05)$; (5) a log-normal distribution with mean zero and unit variance on the log scale, which is also a special case of the Tukey gh distribution with $g = 1$ and $h = 0$; (6) a chi-squared distribution with 3 degrees of freedom, Chisq(3); and (7) a beta distribution with shape parameters $(2, 4)$. All random effects distributions were standardised to have mean zero and unit variance. For each value of c_0 and true random effects distribution, we simulated 500 datasets.

As seen in Figure 1, the true random effects distributions cover of variety of shapes. The normal and $t(3)$ are the only symmetric distributions, while as we transition from the Tukey gh distribution with $g = 0.25$ to $g = 0.5$ to the log-normal distribution, the amount of right skew increases along with the thickness of the right tail. The normal distribution has the thinnest (exponential) tails, while the log normal has the thickest tail followed by the $t(3)$ distribution. The Chisq(3) distribution has a right tail whose thickness lies between the Tukey gh with $g = 0.25$ and Tukey gh with $g = 0.5$. Finally, all but the last three distributions are defined on the entire real line, while the Chisq(3) and log-normal distributions are defined on the positive real line, and the beta distribution is defined on the $[0, 1]$ interval.

For each simulated dataset, we fitted and performed inference on LMMs with the same mean structure as the true model, that is, $\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\Lambda}\boldsymbol{b}_i$, and assuming normally distributed random effects. All model fitting was performed based on direct maximisation of (1) using Nelder–Mead optimisation in `optim` in `R`. We chose this approach to fitting LMMs, as opposed to using the `lme4` package (Bates *et al.*, 2015), say, because it provided more straightforward customisation. As discussed in the succeeding text, for hypothesis testing, we required setting the fixed effect coefficients and standard deviation component at specific non-zero values. Such

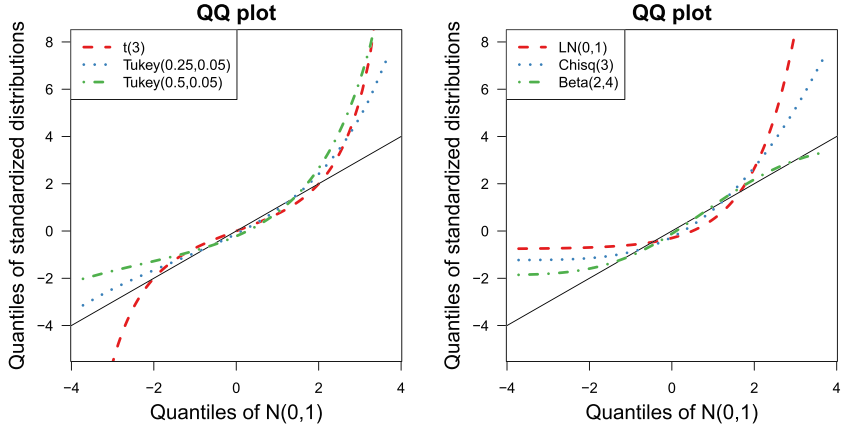

Figure 1. *Quantile–quantile plots of the random effects distributions considered in the simulation. All distributions have been standardised to have mean zero and unit variance. The left panel shows the three distributions that span the real line, and the right panel shows the three distributions whose support is truncated. In both panels, the standard normal distribution is represented by the one-to-one diagonal solid line. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

constraints are not easily implemented in off-the-shelf packages for likelihood-based estimation of mixed models, and so we opted to use a more generic optimisation algorithm. We did however conduct additional simulations to verify that, in the settings where it was possible to use both approaches, `lme4` and Nelder–Mead optimisation produced the same estimates; see also our discussion in the succeeding text for computing confidence intervals.

We consider a variety of measures for assessing performance. For point estimation, we calculated the empirical bias and MSE of the estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, with focus particularly on the standard deviation components. For inference based on hypothesis testing, we performed a standard likelihood ratio test with nominal significance level 5% for three null hypotheses: (1) $H_0 : \beta_1 = K_1$, which is a test of the fixed slope for a covariate that is included as both fixed and random effects; (2) $H_0 : \beta_2 = K_2$, which is a test of the fixed slope for a covariate included as a fixed effect only; and (3) $H_0 : \lambda_2 = K_3$, which is a test of the standard deviation component for a covariate that is included as both fixed and random effects. All likelihood ratio test statistics were compared against a Chisq(1) distribution. We calculated empirical type I errors when K_1, K_2 and K_3 were set to their respective true parameter values and empirical power at two possible values where K_1, K_2 and K_3 were set to two values not equal to their respective true parameter values. Specifically, we considered $K_1 = K_2 = \{0.5, 1, 1.25\}$ and $K_3 = \{0.75c_0, c_0, 1.5c_0\}$, where c_0 is the value of the true standard deviation component for the dataset in question. We also assessed inference based on 95% profile likelihood confidence intervals of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. These intervals were obtained from fitting an LMM using the `lme4` package (noting earlier that `lme4` and Nelder–Mead optimisation produced the same estimates), and then applying the `confint.merMod` function. In summary, profile likelihood confidence intervals are computed based on profile likelihood ratio statistic (where nuisance parameters are replaced by their estimates) and assuming a chi-squared distribution for the statistic with degrees of freedom equal to the number of estimated parameters in the model (Bates *et al.*, 2015). Apart from being the default approach given by `lme4`, using profile likelihood confidence intervals (as opposed to, say, Wald confidence intervals) provides the most appropriate complement to the results obtained for hypothesis testing.

Finally, we considered two numerical measures to assess whether violation of the normality assumption for the random effects distribution can be detected using relatively simple numerical measures. While a common approach taken to diagnose misspecification is to construct normal quantile plots of the predicted random effects, this is not feasible here given the large number of simulation datasets. Furthermore, as Schtzenmeister and Piepho (2012) among others have pointed out, the use of graphical approaches involves a level of subjectivity that applied researchers are often uncomfortable with, and so it is desirable to complement those with numerical measures (see also the recent graphical approach by Efendi *et al.*, 2017). The first measure we consider is a p-value obtained from applying a Shapiro–Wilk test of normality to the predicted random effects. Based on a nominal significance level of 5%, we can then calculate the empirical power at detecting violations in the normality assumption. The second measure we consider is a goodness-of-fit statistic comparing the predicted random effects against a reference random effects distribution, which takes the form $G(\hat{F}, F_0) = \int \{\hat{F}(x) - F_0(x)\}^2 w(x) dF_0(x)$, where $\hat{F}(x)$ and $F_0(x)$ generically denote the empirical cumulative distribution function of the predicted random effects and the reference random effects distribution, respectively. For the weights $w(x)$, we considered: (1) $w(x) = 1$, which corresponds to the Cramér–von Mises criterion, and (2) $w(x) = \{F_0(x)(1 - F_0(x)\}^{-1}$, which corresponds to the Anderson–Darling criterion and weights both tails more heavily. Note that both criteria could also be used to test against normality of the random effects, but our usage is instead as a relative goodness-of-fit statistic. Specifically, for each simulated dataset, we calculated two values of $G(\hat{F}, F_0)$: one where $F_0(x)$ was set to the true random effects distribution

and the other where F_0 was set to the assumed random effects distribution, that is, the standard normal distribution. We then calculated the difference between these two statistics, denoted here as Δ, where $\Delta > 0$ means $G(\hat{F}, F_0)$ is larger when $F_0(x)$ was set to the true random effects distribution. This in turn implies that the predicted random effects more closely followed the assumed random effects distribution. Values of $\Delta > 0$ therefore suggest a 'supernormality' phenomenon, where the predicted random effects more closely follow the assumed rather than the true random effects distribution. Note that if the true random effects are normally distributed, then $\Delta = 0$. It is important to keep in mind that what constitutes a large difference depends on both the distribution $F_0(x)$ and the weights $w(x)$ used in the goodness-of-fit statistic.

5 Results

We first summarise the results for fixed effects inference, before proceeding in detail to random effects estimation and inference. The detailed results for fixed effects inference are presented in the Supporting Information. For LMMs, overall results showed that maximum likelihood estimation and inference on the fixed effects slopes, irrespective of whether the corresponding covariate is also included as random slope or not, were largely robust to the misspecification. There was little difference in the biases and MSE of the estimates of $\boldsymbol{\beta}$ for all seven true random effects distributions and true standard deviation c_0 considered. As expected, the bias and MSE were smallest for estimates of $\hat{\beta}_2$, which was included as fixed effect only. Increasing the number of clusters improved point estimation, for example, the MSEs of the fixed effects coefficients approximately halved when n was doubled from 50 to 100, but for fixed n, increasing cluster size had little impact on point estimation. The robustness in the estimation of the fixed effects generally carried over to hypothesis testing and confidence interval coverage: in agreement with Result 1, the type I errors for testing both fixed effects were close to the nominal level, while the empirical coverage probability of the profile confidence intervals hovered around the nominated 95% level. The notable exception to this was hypothesis testing of β_1 and coverage probabilities of β_0 and β_1, when the true random effects distribution was log normal. In this setting, the type I errors were not only inflated and increased as the true standard deviation c_0 increased but also tended to be worse when the cluster size increased. Analogously, the empirical coverage probability dropped noticeably below the nominated 95% level as the true standard deviation c_0 increased. For hypothesis testing, the power to detect a non-zero fixed effect whose covariate was also included as a random effect decreased with increasing c_0. Finally, similar to the point estimation results (and consistent with Result 2), increasing the number of clusters improved power and type I error performance, while increasing cluster size had little impact on significance testing, aside from in the log-normal distribution case already discussed.

5.1 Estimation of Variance Components

Point estimation of the standard deviation components was sensitive to misspecification of the random effects distribution and the size of the true standard deviation c_0 (Figure 2). Estimates of both λ_1 and λ_2 tended to be negatively biased, with the bias most extreme for the three heavy-tailed true random effects distributions [standardised log normal, $t(3)$, and Tukey gh distribution with $(g = 0.5, h = 0.05)$]. It should be pointed out that on a relative scale, the relative biases, defined as the bias divided by the true value of the standard deviation, range up to a maximum of about -12.2% in our simulation design.

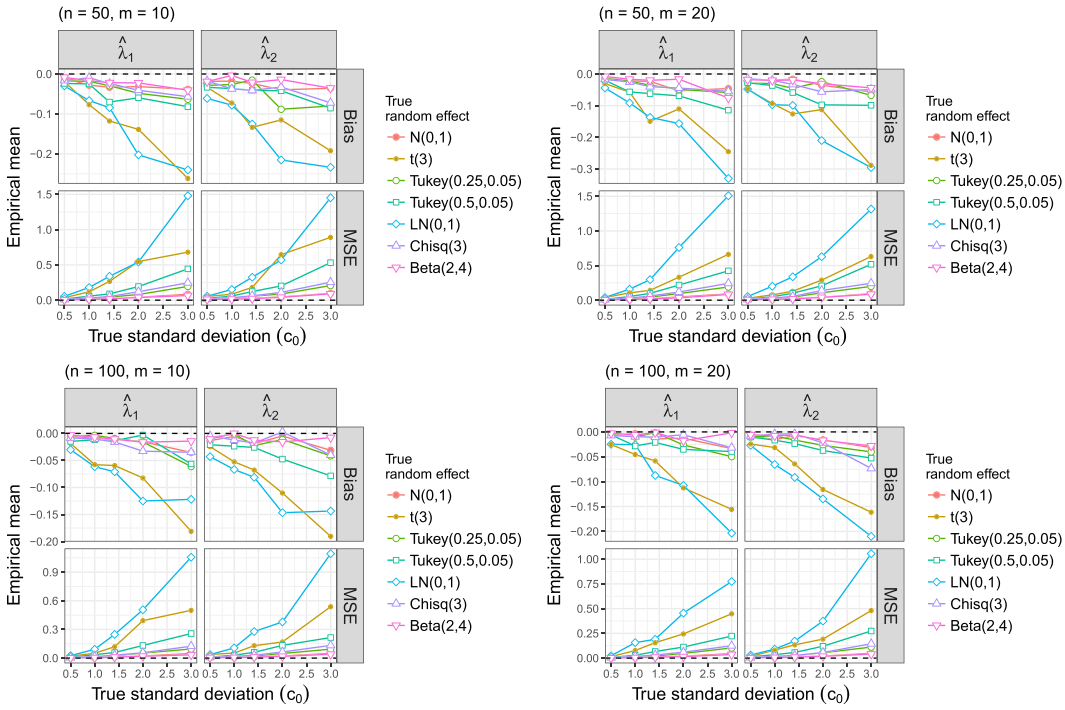

Figure 2. *Empirical bias (top row) and mean squared error (MSE) (bottom row) of the point estimates for standard deviation components* $\boldsymbol{\lambda}$ *against difference values of the true standard deviation and under different true random effects distributions. Simulation results are presented for combinations of number of clusters* $\boldsymbol{n} = \{50, 100\}$ *and cluster size* $\boldsymbol{m} = \{10, 20\}$. *[Colour figure can be viewed at* wileyonlinelibrary.com*]*

According to McCulloch and Neuhaus (2011a), an asymptotic relative bias on the order of $\pm 15\%$ or less amounts to 'little effect on estimation of the random effects variance'. We disagree with this assessment however, especially because across all true random effects distributions and values of c_0 considered in our simulation, the maximum absolute relative biases incurred by any of the fixed effects were less than 5%. Comparatively speaking, the bias incurred by the point estimates of the standard deviation components is rather high.

We also observed some negative bias at $n = 50$ when the random effects distribution was correctly specified, that is, normally distributed. This however can be attributed to the use of maximum likelihood estimation for LMMs, which is well known to produce finite sample bias and to underestimate the variance components even when the model is correctly specified. Indeed, when we reran all the simulations using restricted maximum likelihood estimation (Patterson & Thompson, 1971), the bias was much closer to zero (not shown) when the random effects distribution was correctly specified.

At first glance, the direction of the biases observed under random effects misspecification seems counterintuitive: if the true random effects distribution is heavy tailed but we assume normality, then it seems more logical for the variance components to be overestimated in order to compensate for this. To understand why the biases are actually negative, first note that the estimates of variance components are typically based on the between-cluster MSE. This is strictly true if we estimated the LMM using analysis of variance, and in the case of independent random effects and balanced data (as in our simulation design), this also holds true when the variance components are estimated using restricted maximum likelihood estimation (Jiang, 2007). Next, consider what happens when constructing a between-cluster MSE for a standard normal versus a standardised heavily tailed random effects distribution: both will have the same expected

value of one, but because the latter has a heavier right tail, then the bulk of its distribution will actually lie below that of the former.

As an example, consider the quantity $Q = n^{-1}\sum_{i=1}^{n} Z_i^2$, where the Z_is are independent standard normal versus independent standardised $t(3)$ distributions. For the former, $Z_i^2 \sim$ Chisq(1), and thus, $Q \sim n^{-1}$Chisq(n). For the latter, $Z_i^2 = 3^{-1}F(1,3)$, where $F(1,3)$ is the F-distribution with degrees of freedom equal to 1 and 3, and hence, Q is equal to the sum of n-independent $F(1,3)$ distributions divided by $3n$. Compared with a Chisq(1), the $F(1,3)$ distribution has a heavier tail and the bulk of its distribution lies below that of the Chisq(1). Critically, because in finite samples, for example, $n = 50$, the quantity Q is more strongly influenced by the bulk rather than the tail of the distribution of the Z_is, this causes the value of Q and generally the between-cluster MSE to be underestimated. Of course, as the number of clusters $n \to \infty$, the law of large numbers implies that the estimates of the variance component are asymptotically unbiased and consistent, and indeed, this is what was seen in the simulations (see Figure 2). Thus, this bias is a finite sample phenomenon, and importantly, it is the sensitivity of this bias to random effects misspecification that presents a concern.

The strong biases suffered under true heavy-tailed random effects distributions carried over to the MSEs, where there was considerable sensitivity to random effects misspecification and the value of c_0 (Figure 2). For the case of the standardised log normal, $t(3)$, and Tukey gh with ($g = 0.5, h = 0.05$) random effects distribution, comparing the MSEs to the squared biases suggests that there was substantial variability in the estimates of standard deviation components across simulated datasets. This is a cause for concern not only because such variability feeds into inference, as we shall see shortly, but also because estimates of variance components are often used in quantifying the proportion of variance explained among other interpreted measures in mixed model applications in ecology and genetics (as reviewed in Section 1). On a relative scale, across the true random effects distributions and four combinations of n and m tested, the relative root MSE, defined as the square root of the MSE divided by the true value of the standard deviation, averaged 22% and ranged up to 47%. By comparison, the maximum relative root MSE across all fixed effects averaged 14%, although it did range up to 46% (see the Supporting Information).

Both bias and MSE of the variance component estimates decreased as the number of clusters increases. For fixed n, doubling the cluster size did not produce any noticeable change in point estimation performance. This result is not overly surprising and reinforces previous theoretical results concerning mixed models in general that $m \to \infty$ alone is insufficient for achieving consistency of the maximum likelihood estimates of the variance components (Nie, 2007).

5.2 Hypothesis Testing and Confidence Intervals for Variance Components

In agreement with Result 3, likelihood ratio hypothesis testing of truly non-zero standard deviation parameters is very sensitive to misspecification of the random effects distribution (Figure 3). With the exception of the normal (such that true and assumed random effects distribution matched) and Beta(2, 4) distributions (which is doubly truncated), misspecification of the random effects distribution leads to substantially inflated type I errors with the heavier tailed distributions suffering worse inflation. The type I error also tended to be worse as c_0 increased, which complements the worsening biases and MSEs seen in point estimation (Figure 2). For fixed n, increasing the cluster size has little impact on type I errors under misspecification. Furthermore, in agreement with Result 3, doubling the number of clusters n from 50 to 100 made negligible difference to the type I error when testing truly non-zero variance components. In fact, when the true random effects were Beta(2, 4) distributed, there was evidence that the test became conservative.

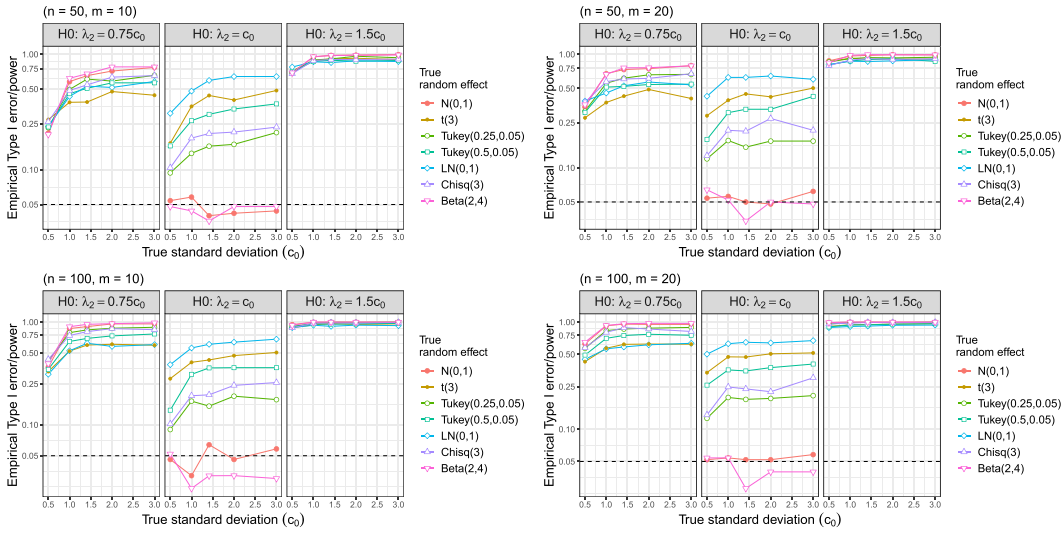

Figure 3. *Empirical type I error (middle panel) and power (left and right panel) for testing hypotheses concerning* λ_2 *against difference values of the true standard deviation and under different true random effects distributions. Simulation results are presented for combinations of number of clusters* $n = \{50, 100\}$ *and cluster size* $m = \{10, 20\}$. *[Colour figure can be viewed at* wileyonlinelibrary.com*]*

In terms of power, it is interesting to note that misspecification of the random effects distribution had mostly a detrimental consequence on the power to detect departures from the true variance component values (Figure 2). Increasing the number of clusters leads to considerable improvement in power for all true random effects distributions, while doubling the cluster size also led to smaller improvements. We caution into reading these results too deeply, if at all, however, given the standard likelihood ratio test has been shown earlier to be invalid, that is, the type I error is not (asymptotically) at the nominal level, for assessing non-zero standard deviation components.

The results obtained for the empirical coverage probability of profile likelihood confidence intervals for standard deviation components was consistent with those obtained earlier for hypothesis testing. Coverage probability was very sensitive to misspecification of the random effects distribution (Figure 4), with all but the normal and Beta$(2, 4)$ distributions producing coverage noticeably below the nominal 95% level. The heavier tailed distributions in particular suffered the worst under coverage, dropping to as low as around 30%. Increasing both n and m had little impact on coverage probability under random effects misspecification, and in fact when the true random effects were Beta$(2, 4)$ distributed, there was evidence of over coverage, which was consistent with evidence of the test being conservative in Figure 3. Although not performed, we anticipate that such under coverage when the random effects distribution is misspecified will also be seen for other types of confidence intervals, notably for Wald-type confidence intervals.

5.3 Diagnosing Misspecification

A simple Shapiro–Wilk test on the predicted random effects performed reasonably well at detecting departures from the normality assumption (Table 1), although there was some sensitivity to the true random effects distribution. Power tended to be highest for skewed distributions with tails heavier than the normal, that is, the log normal and Chisq(3). As expected, the power

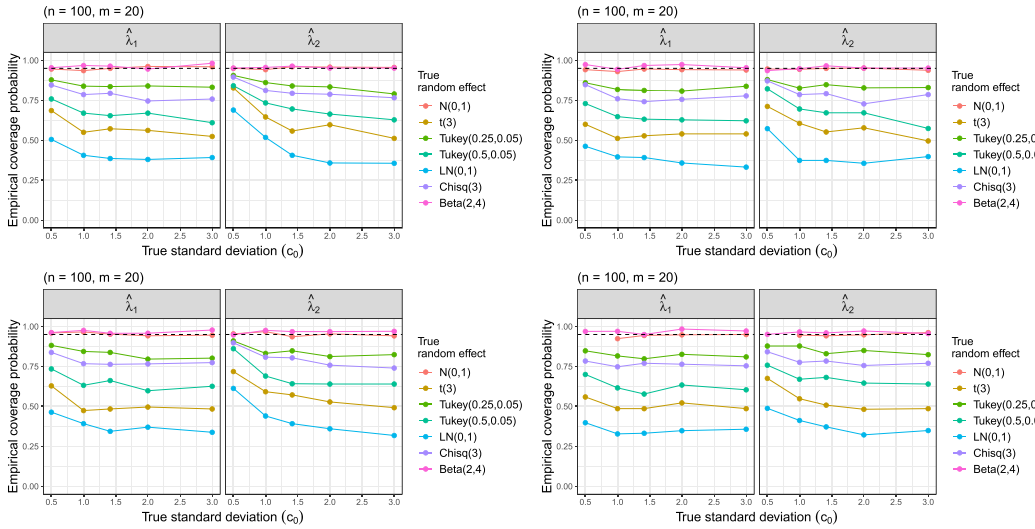

Figure 4. *Empirical coverage probability of 95% profile likelihood confidence intervals for* λ_1 *and* λ_2 *against difference values of the true standard deviation and under different true random effects distributions. Simulation results are presented for combinations of number of clusters* $n = \{50, 100\}$ *and cluster size* $m = \{10, 20\}$. *[Colour figure can be viewed at* wileyonlinelibrary.com*]*

increases as the true standard deviation becomes larger. The power to detect misspecification was also substantially higher for the random intercept b_{i0} compared with the slope b_{i1}. There were also substantial improvements in power when the number of clusters and/or cluster size was increased. This makes sense given the ability to detect departures from normality is expected to improve both when there are more predicted random effects (larger sample size) and when we have more information to predict each random effect (larger cluster size).

Complementing the reasonable performance of the more formal Shapiro–Wilk test were boxplots of the differences in goodness-of-fit statistics, which overall presented evidence that the predicted random effects for both random intercept and slope tended to follow the true random effects more closely than the assumed normal distribution (Figure 5 and the figures in the Supporting Information). This was especially the case for the Chisq(3), the two Tukey gh distributions and the Beta$(2, 4)$ random effects distributions, where across the four combinations of n and m tested, 75% or more of the differences in goodness-of-fit statistics tended to be less than zero for both weighting schemes used when $c_0 > \sqrt{2}$. The log-normal distribution presented a rather interesting result, with the results being almost mirror opposites for the two weighting schemes. This arises because in the log-normal case, the shrinkage factor (see McCulloch & Neuhaus, 2011b) causes the predicted random effects to be considerably less right skewed compared with their true values. When more weight is put on the tails of the distribution, then as in the Anderson–Darling criterion (right panels in Figure 5 and the figures in the Supporting Information), the predicted random effects present evidence of supernormality. For an equal weighting scheme, however, as in the Cramér–von Mises criterion (left panels in Figure 5 and the figures in the Supporting Information), the asymmetry of the predicted random effects becomes the dominate characteristic instead, and thus, the predicted random effects present departure from the assumed normality. Finally, for the $t(3)$ random effects distribution, the differences in goodness-of-fit statistic were largely scattered around zero for both the predicted random intercept and slope when $n = 50$, irrespective of the weighting scheme and the true standard deviation. This was consistent with the intermediate but not strong power at being to diagnose misspecification from the Shapiro–Wilk test at $n = 50$ (Table 1). Similar

Table 1. *Empirical power of the Shapiro–Wilk test applied to the predicted random intercept ($\hat{b}_{i0}$) and slope ($\hat{b}_{i1}$), for detecting violations of the normality assumption for the random effects distributions.*

		$(n = 50, m = 10)$		$(n = 50, m = 20)$		$(n = 100, m = 10)$		$(n = 100, m = 20)$	
True random effect	c_0	$\hat{b}_{i0}$	$\hat{b}_{i1}$	$\hat{b}_{i0}$	$\hat{b}_{i1}$	$\hat{b}_{i0}$	$\hat{b}_{i1}$	$\hat{b}_{i0}$	$\hat{b}_{i1}$
	0.5	0.052	0.040	0.044	0.060	0.052	0.056	0.056	0.060
	1	0.080	0.036	0.038	0.048	0.034	0.068	0.052	0.046
$\mathcal{N}(0, 1)$	$\sqrt{2}$	0.048	0.048	0.044	0.040	0.046	0.058	0.052	0.056
	2	0.050	0.052	0.038	0.056	0.050	0.048	0.050	0.046
	3	0.054	0.048	0.060	0.038	0.060	0.050	0.050	0.048
	0.5	0.412	0.242	0.478	0.304	0.610	0.378	0.728	0.516
	1	0.562	0.464	0.540	0.496	0.780	0.714	0.830	0.734
$t(3)$	$\sqrt{2}$	0.574	0.530	0.562	0.558	0.838	0.774	0.866	0.806
	2	0.598	0.578	0.642	0.654	0.866	0.814	0.842	0.808
	3	0.604	0.614	0.638	0.658	0.876	0.846	0.866	0.852
	0.5	0.266	0.162	0.308	0.188	0.456	0.252	0.590	0.384
	1	0.445	0.331	0.470	0.372	0.682	0.570	0.728	0.634
Tukey$(0.25, 0.05)$	$\sqrt{2}$	0.432	0.414	0.486	0.394	0.696	0.640	0.724	0.712
	2	0.474	0.416	0.474	0.462	0.726	0.694	0.784	0.756
	3	0.470	0.482	0.480	0.460	0.732	0.762	0.776	0.752
	0.5	0.517	0.337	0.720	0.492	0.802	0.524	0.944	0.734
	1	0.816	0.668	0.820	0.744	0.980	0.896	0.994	0.990
Tukey$(0.5, 0.05)$	$\sqrt{2}$	0.854	0.786	0.860	0.808	0.990	0.944	0.998	0.990
	2	0.888	0.850	0.906	0.860	0.992	0.974	0.996	0.992
	3	0.902	0.880	0.908	0.884	0.996	0.990	0.992	0.992
	0.5	0.732	0.468	0.881	0.608	0.930	0.666	0.996	0.873
	1	0.950	0.856	0.992	0.924	1	0.984	1	0.998
LN$(0, 1)$	$\sqrt{2}$	0.984	0.944	0.994	0.974	1	0.996	1	1
	2	0.994	0.968	1	0.996	1	1	1	1
	3	1	0.998	1	1	1	1	1	1
	0.5	0.556	0.296	0.712	0.420	0.842	0.494	0.964	0.744
	1	0.890	0.680	0.918	0.818	0.990	0.943	0.998	0.988
Chisq(3)	$\sqrt{2}$	0.964	0.842	0.964	0.914	1	0.988	1	1
	2	0.982	0.914	0.986	0.958	1	0.998	1	1
	3	0.982	0.972	0.990	0.972	1	0.998	1	1
	0.5	0.108	0.056	0.176	0.076	0.202	0.074	0.298	0.116
	1	0.198	0.148	0.278	0.200	0.528	0.308	0.618	0.422
Beta$(2, 4)$	$\sqrt{2}$	0.268	0.230	0.296	0.274	0.613	0.462	0.680	0.562
	2	0.316	0.282	0.336	0.286	0.676	0.600	0.768	0.666
	3	0.328	0.312	0.348	0.366	0.732	0.636	0.768	0.702

Empirical power was calculated as the proportion of (500) datasets where the p-value from the test was less than 0.05.

to the log-normal distribution case, the shrinkage factor caused the predicted random effects to be considerably less heavy tailed than their true values. Together with the symmetry of both the normal and t-distributions, then there was little evidence to conclusively favour either the assumed or true random effects distribution. Performance improved considerably when the number of clusters was doubled to 100. More generally, the magnitudes of the differences (whether they tended to be positive or negative) increased with a larger number of clusters n, while increasing cluster size m had less impact.

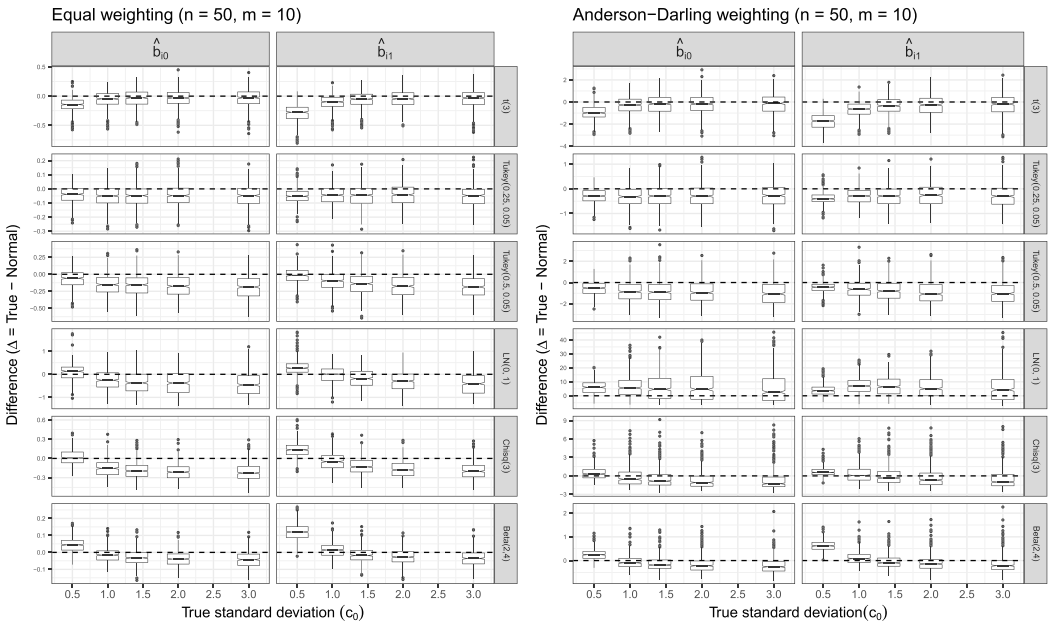

Figure 5. *Boxplots of differences in goodness-of-fit statistic* Δ *for the predicted random intercept* $\hat{b}_{i0}$ *and slope* $\hat{b}_{i1}$ *with equal (left) and Anderson–Darling (right) weighting schemes at* $(n = 50, m = 10)$. *A difference* $\Delta > 0$ *implies that the predicted random effects more closely followed the assumed random effects distribution than the true random distribution.*

6 Application to High School and Beyond Survey

To illustrate the effects of random effects misspecification on a real data analysis, we fitted LMMs to the 1982 HSB survey. The full data are available from the package `mlmRev` (Bates *et al.*, 2014), although for illustrative purposes, we consider a subsample of 4 460 students nested within $n = 100$ schools (see Zwick & Sklar, 2005; Hui *et al.*, 2019, for other previous applications of the HSB survey).

The dataset comprised a standardised, continuous measure of mathematics achievement as the response, along with one school-level covariate (sector, which is equal to 0 for public school and 1 for Catholic school) and three student-level covariates (gender, which is equal to 0 for female and 1 for male; race, which is equal to 1 if the student belonged to minority race and 0 otherwise; and SES, which is a continuous social economic status score reflecting parental education, occupation and income). The number of students recorded within each school, that is, cluster size, ranged from $m = 14$ up to $m = 67$.

We began by fitting an LMM including all four covariates as fixed effects ($p_f = 5$ including the fixed effect intercept), a random intercept for school, and independent random slopes for school for each of the three student-level covariates ($p_r = 4$). Results indicate that on average across all schools, mathematics achievement tended to be better for public schools and students who were male, of a non-minority race, and had higher social economic status (Table 2 first column). Based on the estimates of the standard deviation components, there was also evidence of considerable heterogeneity between schools in the 'baseline' mathematics achievement scores, as well as the impact of gender, race and to a much lesser extent SES, on these scores. Both quantile–quantile plots and Shapiro–Wilk tests for each of the four sets of predicted random effects suggest no strong evidence of departures from the assumption of normality except for race, which showed evidence of being right skewed (see the Supporting Information, and

Table 2. *Estimates from fitting linear mixed models to either the original HSB survey or modified versions of the survey where the (standardised) random effects generated several non-normal distributions.*

Estimates	Original data	$\mathcal{N}(0,1)$	Tukey$(0.25, 0.05)$	LN$(0,1)$	Beta$(2,4)$
$\hat{\beta}_0$	14.159 (13.587, 14.731)	13.725 (13.187, 14.262)	13.612 (13.061, 14.158)	13.911 (13.406, 14.415)	14.337 (13.770, 14.904)
$\hat{\beta}_{\text{Sector}}$	−2.172 (−2.901, −1.450)	−1.952 (−2.629, −1.265)	−1.673 (−2.366, −0.979)	−1.823 (−2.451, −1.192)	−2.364 (−3.074, −1.656)
$\hat{\beta}_{\text{Gender}}$	1.128 (0.647, 1.611)	1.134 (0.928, 1.944)	1.055 (0.602, 1.507)	1.198 (0.725, 1.674)	1.022 (0.568, 1.478)
$\hat{\beta}_{\text{Race}}$	−3.100 (−3.665, −2.542)	−2.672 (−3.239, −2.103)	−3.648 (−4.236, −3.061)	−3.064 (−3.570, −2.560)	−2.914 (−3.432, −2.381)
$\hat{\beta}_{\text{SES}}$	2.065 (1.781, 2.349)	2.032 (1.760, 2.305)	2.018 (1.748, 2.290)	2.084 (1.813, 2.354)	2.170 (1.897, 2.444)
$\hat{\lambda}_0$	1.400 (1.054, 1.775)	1.229 (0.909, 1.583)	1.328 (1.029, 1.665)	1.121 (0.842, 1.433)	1.428 (1.136, 1.761)
$\hat{\lambda}_{\text{Gender}}$	1.226 (0.634, 1.773)	1.387 (0.864, 1.915)	0.927 (0, 1.490)	1.144 (0.600, 1.651)	0.896 (0, 1.473)
$\hat{\lambda}_{\text{Race}}$	0.929 (0, 1.709)	0.910 (0, 1.645)	1.174 (0.245, 1.879)	<0.01 (0, 1.018)	0.167 (0, 1.293)
$\hat{\lambda}_{\text{SES}}$	0.401 (0, 0.884)	<0.01 (0, 0.708)	0.038 (0, 0.773)	<0.01 (0, 0.594)	<0.01 (0, 0.716)

Immediately below each estimate is the corresponding 95% profile confidence interval.

recall from Section 5.3 that the Shapiro–Wilk test seemed to work relatively well at diagnosing misspecification).

To assess how sensitive the estimates and inference for the HSB survey was to different random effects distributions, we generated seven new sets of random effects $\boldsymbol{b}_i = (b_{i0}, b_{i\text{Gender}}, b_{i,\text{Race}}, b_{i\text{SES}})^\top$ for $i = 1, \ldots, 100$ by simulating from the seven standardised distributions considered in Section 4. Conditional on each set of random effects, and treating the parameter estimates from model fitted to the original dataset as the true parameters, we then simulated responses from an LMM. Finally, we fitted the same LMM (assuming normally distributed random effects) as was performed for the original HSB survey to each set of random effects and compared estimates and 95% profile confidence intervals of the fixed effects coefficients and standard deviation components. In the second to fourth columns of Table 2, we present results for four of the seven sets, while the full results are presented in the Supporting Information.

The fixed effects estimates of gender and SES did not vary much across the different sets of random effects distributions, while the estimates for sector and race were quite sensitive (Table 2 second to fifth columns). This sensitivity may be largely a result of natural variation in simulation (given we only simulate one dataset here for each set of random effect), meaning we cannot expect the robustness of fixed effects inference observed in Section 5 to be fully realised here. Nevertheless, it was surprising to see how much the fixed effects estimates and location of the confidence intervals changed across different true random effects distributions, particularly given the size of the dataset with 100 clusters and more than 4 400 observations in total. Consistent with Sections 5.1 and 5.2, estimates of all the standard deviation components and the location/width of the corresponding profile confidence intervals were sensitive to true random effects distribution. This was exemplified by the standard deviation component for race, where the estimates ranged from <0.01 to 1.174 and the confidence interval widths ranged from

approximately 1 to 1.6, and for SES, where the estimates of λ_{SES} were all close to zero even though the true value of 0.401 and the confidence interval widths ranged from 0.6 to 0.9.

7 Discussion

In the summary of their review paper, McCulloch and Neuhaus (2011a) conclude that a wide array of inferences in mixed models, including 'estimation of the random effects variance', is robust to misspecification of the random effects distribution. While we commend them for providing a thorough and engaging review of the conflicting literature on misspecification, which was summarised to a lesser extent in Section 1, we worry that their overall message, which is skewed heavily by a strong interest in fixed effects inference, has led many applied researchers to be naively confident in the assumption of normal random effects distribution and its robustness. Further exacerbating this problem is the fact that the assumption of normality dominates software for fitting mixed models, for example, popular `R` packages such as `lme4`, `glmmPQL` (Venables & Ripley, 2002) and `MCMCglmm` (Hadfield, 2010), only permit normal random effects. By contrast, the bulk of methodological research into mixed models with other random effects distributions (e.g. Zhang & Davidian, 2001; Gu & Ma, 2005; Papageorgiou & Hinde, 2012) has gathered less attention from applied researchers.

In this article, we have presented theoretical and empirical evidence that random effects inference in LMMs can be severely impacted by misspecification of the random effects distribution. Estimates of the standard deviation components are subject to considerable finite sample negative bias, standard likelihood ratio testing (when the standard deviation components is truly non-zero) does not produce a valid test and presents substantially inflated type I errors if the true random effects distribution is heavy tailed and profile likelihood confidence intervals can exhibit major under coverage. In fact, one interesting trend observed across all values of n and m tested was that the heavier the tail of the true random effects distribution, the more severe the type I error inflation when testing variance components. On the opposite end of the spectrum, the beta random effects distribution, which is doubly truncated, showed evidence of being an overly conservative test. Furthermore, we cannot rely on the robustness brought about by having a larger number of clusters n in this situation, as we have shown that the standard likelihood ratio test is *asymptotically* invalid when testing truly non-zero variance components. That is, unlike point estimation and hypothesis testing of fixed effects, and point estimation of variance components, the consequences of random effects misspecification on inference of non-zero variance components do *not* vanish asymptotically.

Fortunately, even when the number of clusters and cluster size is relatively small, simple goodness-of-fit statistics and tests of normality do have reasonable power to diagnose misspecification, although there is again some sensitivity to the shape of the true random effects distribution and weights used in the test statistic. Our simulations also showed that increasing cluster size had negligible impact on random effects inference compared with increasing the number of clusters. We point out that increasing cluster size (at the same time as the number of clusters) is a common and realistic assumption made when studying inference for mixed models for independent clustered data (e.g. Hui *et al.*, 2017), yet exploring how random effects misspecification specifically is affected by increasing cluster size relative to increasing the number of clusters has received comparably little attention (see Neuhaus & McCulloch, 2011; Zhang *et al.*, 2016, for some exceptions).

Our simulations have several implications. From a computational viewpoint, more software needs to be developed to give applied researchers the capacity to fit mixed models with random effects distributions beyond that of the normal, in conjunction with an array of diagnostic

tools for diagnosing misspecification. Analogous to selecting prior distributions for Bayesian estimation and inference, an increased choice of random effects distributions both offers a form of sensitivity analysis and allows the incorporation of prior scientific knowledge on the between-cluster variability. Building on this, we also advocate that more theoretical research and software needs to be developed for robust inference of mixed models, especially robust hypothesis testing for random effects inference (see, for instance, very early work by Kent, 1982, on robust likelihood ratio tests).

Although the focus of this paper was on LMMs, an important avenue of future research is to explore how random effects inference is impacted by random effects misspecification in non-linear and generalised linear mixed-effects models where the assumption of normality on the response is relaxed. We conjecture that the theoretical and empirical results presented here will also hold when the assumption of normality is relaxed and replaced with more general conditions on the marginal moments of the response, especially because the derivations of Results 1 and 3 rely more on moment rather than distributional results (see also Richardson & Welsh, 1994). Covering the broader case of the responses from the exponential family is of considerable interest, given random effects inference tends to often be applied to generalised linear mixed-effects models more so than LMMs; see, for example, and Nakagawa & Schielzeth (2013) but also more broadly for generalised linear latent variables models where the multivariate random effects (latent variables) are often of interest (e.g. Hui, 2017).

Finally, while the focus of this article has been on the ability to perform inference under random effects misspecification, it is important to try and avoid unnecessary random effects in the model to begin with, that is, select only truly important random effects. One approach to performing such selection is hypothesis testing, and as discussed in Section 1, the standard likelihood ratio test for assessing zero variance components is robust to random effects misspecification (although they tend to be overly conservative, Stram & Lee, 1994). More broadly, random effects selection in mixed models remains an active topic in statistical research, with methods ranging from hypothesis testing including simulation-based tests (Greven *et al.*, 2008; Giampaoli & Singer, 2009; Drikvandi *et al.*, 2013; Hui *et al.*, 2019) to penalised likelihood methods (Pan & Huang, 2014; Hui *et al.*, 2017). However, the degree in which such selection may be impacted by misspecification of the distribution of other random effects included the model, which consequently means that the marginal log likelihood used as the basis for selection is misspecified, largely unexplored and a topic worthy of future research.

Acknowledgements

This research was partly supported by the Australian Research Council discovery project grant DP180100836. Thanks to Emi Tanaka and Robert Clark for their useful discussions.

References

Agresti, A., Caffo, B. & Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Comput. Stat. Data Anal.*, **47**, 639–653.

Bates, D., Maechler, M. & Bolker, B. (2014). *mlmRev: examples from multilevel modelling software review*. R package version 1.0-6.

Bates, D., Mchler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, **67**, 1–48.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. & White, J. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol*, **24**, 127–135.

Butler, S. M. & Louis, T. A. (1992). Random effects models with non-parametric priors. *Stat. Med.*, **11**, 198–12000.

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V. & Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecol. Appl.*, **19**, 553–570.
Drikvandi, R., Verbeke, G., Khodadadi, A. & Nia, V. P. (2013). Testing multiple variance components in linear mixed-effects models. *Biostatistics*, **14**, 144–159.
Drikvandi, R., Verbeke, G. & Molenberghs, G. (2017). Diagnosing misspecification of the random-effects distribution in mixed models. *Biometrics*, **73**, 63–71.
Efendi, A., Drikvandi, R., Verbeke, G. & Molenberghs, G. (2017). A goodness-of-fit test for the random-effects distribution in mixed models. *Stat. Methods Med. Res.*, **26**, 970–783.
Gaecki, A. & Burzykowski, T. (2013). *Linear Mixed-effects Models Using R: A Step-by-Step Approach.* New York: Springer.
Giampaoli, V. & Singer, J. M. (2009). Likelihood ratio tests for variance components in linear mixed models. *J. Stat. Plann. Infer.*, **139**, 1435–1448.
Greven, S., Crainiceanu, C. M., Kuchenhoff, H. & Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *J. Comput. Graph. Stat.*, **17**, 870–891.
Grilli, L. & Rampichini, C. (2015). Specification of random effects in multilevel models: a review. *Qual. Quant.*, **49**, 967–976.
Gu, C. & Ma, P. (2005). Generalized nonparametric mixed-effect models: computation and smoothing parameter selection. *J. Comput. Graph. Stat.*, **14**, 485–504.
Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.*, **33**, 1–22.
Heagerty, P. J. & Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, **88**, 973–985.
Hui, F. K. C. (2017). Model-based simultaneous clustering and ordination of multivariate abundance data in ecology. *Comput. Stat. Data Anal.*, **105**, 1–10.
Hui, F. K. C., Müller, S. & Welsh, A. H. (2017). Joint selection in mixed models using regularized PQL. *J. Am. Stat. Assoc.*, **112**, 1323–1333.
Hui, F. K. C., Müller, S. & Welsh, A. H. (2019). Testing random effects in linear mixed models: another look at the F-test (with discussion). *Aust. New Zealand J. Stat.*, **61**, 61–84.
Ives, A. R. & Helmus, M. R. (2011). Generalized linear mixed models for phylogenetic analyses of community structure. *Ecol. Monog.*, **81**, 511–525.
Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications.* Spring, New York: Springer.
Jiang, J. & Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, **15**, 1.
Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, **69**, 19–27.
Lee, Y., Nelder, J.A. & Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood.* Boca Raton, Florida: CRC Press.
Li, D., Ives, A. R. & Waller, D. M. (2017). Can functional traits explain phylogenetic signal in the composition of a plant community? *New Phytologist*, **214**, 607–618.
Litire, S., Alonso, A. & Molenberghs, G. (2007). Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics*, **63**, 1038–1044.
Litire, S., Alonso, A. & Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Stat. Med.*, **27**, 3125–3144.
Litire, S., Alonso, A. & Molenberghs, G. (2011). Rejoinder to "A note on type II error under random effects misspecification in generalized linear mixed models". *Biometrics*, **67**, 656–660.
Magder, L. S. & Zeger, S. L. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *J. Am. Stat. Assoc.*, **91**, 1141–1151.
Martin, J. G., Nussey, D. H., Wilson, A. J. & Reale, D. (2011). Measuring individual differences in reaction norms in field and experimental studies: a power analysis of random regression models. *Methods Ecol. Evol.*, **2**, 362–374.
McCulloch, C. E. & Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Stat. Sci.*, **26**, 388–402.
McCulloch, C. E. & Neuhaus, J. M. (2011). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*, **67**, 270–279.
Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol. Evol.*, **4**, 133-142.
Neuhaus, J. M. & McCulloch, C. E. (2011). Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika*, **98**, 147–162.
Neuhaus, J. M., McCulloch, C. E. & Boylan, R. (2011). A note on type II error under random effects misspecification in generalized linear mixed models. *Biometrics*, **67**, 654–656.

Neuhaus, J. M., McCulloch, C. E. & Boylan, R. (2013). Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. *Stat. Med.*, **32**, 2419–2429.
Nie, L. (2007). Convergence rate of MLE in generalized linear and nonlinear mixed-effects models: theory and applications. *J. Stat. Plann. Infer.*, **137**, 1787–1804.
Pan, J. & Huang, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Stat. Comput.*, **24**, 725–738.
Papageorgiou, G. & Hinde, J. (2012). Multivariate generalized linear mixed models with semi-nonparametric and smooth nonparametric random effects densities. *Stat. Comput.*, **22**(1), 79–92.
Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.
Pinheiro, J. & Bates, D. (2006). *Mixed-effects Models in S and S-PLUS.* New York: Springer.
Richardson, A. & Welsh, A. H. (1994). Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models. *Aust. New Zealand J. Stat.*, **36**, 31–43.
Schielzeth, H. & Nakagawa, S. (2013). Nested by design: model fitting and interpretation in a mixed model era. *Methods Ecol. Evol.*, **4**, 14–24.
Schützenmeister, A. & Piepho, H.-P. (2012). Residual analysis of linear mixed models using a simulation approach. *Comput. Stat. Data Anal.*, **56**(6), 1405–1416.
Smith, A. B., Ganesalingam, A., Kuchel, H. & Cullis, B. (2015). Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theor. Appl. Gen.*, **128**, 55–72.
Stram, D. O. & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171–1177.
Tashakkori, A. & Teddlie, C. (2010). *SAGE Handbook of Mixed Methods in Social & Behavioral Research.* Thousand Oaks, California: SAGE Publications.
Van der Vaart, A. W. (2000). *Asymptotic Statistics*, Vol. 3 Cambridge: Cambridge University Press.
Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S.* New York: Springer.
Verbeke, G. & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Comput. Stat. Data Anal.*, **23**, 541–556.
Verbeke, G. & Molenberghs, G. (2013). The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models. *Biostatistics*, **14**, 477–490.
Woodard, D., Love, T., Thurston, S., Ruppert, D., Sathyanarayana, S. & Swan, S. (2013). Latent factor regression models for grouped outcomes. *Biometrics*, **69**, 785–794.
Zhang, D. & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, **57**, 795–802.
Zhang, B., Liu, W., Zhang, H., Chen, Q. & Zhang, Z. (2016). A note on misspecification in joint modeling of correlated data with informative cluster sizes. *J. Stat. Plann. Infer.*, **170**, 46–63.
Zwick, R. & Sklar, J. C. (2005). Predicting college grades and degree completion using high school grades and SAT scores: the role of student ethnicity and first language. *Am. Educ. Res. J.*, **42**, 439–464.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

[*Received April 2018, accepted March 2020*]